

22 Data Reconciliation and Software Methods for Bias Detection

MIGUEL J. BAGAJEWICZ AND SR DERRICK K. ROLLINS

INTRODUCTION

Data reconciliation is a term used for the problem of adjusting plant measurements containing errors (e.g., flow rates, temperatures, pressures, and concentrations), so that they conform to a certain chosen model, usually conservation laws.

Consider f to be the $q \times 1$ vector of all the true q flow rates in a plant and C $q \times p$ the connectivity (or incidence) matrix, such that $Cf=0$ represents all the steady-state material balances in all p units. Consider also a subset of flow rates that are measured (given by the $q_M \times 1$ vector f_M). Let f_M^* be the estimator of f . Then f_M^* is obtained solving the steady-state data reconciliation problem, which in its simplest version (the one used in commercial software) is the following:

$$\left. \begin{array}{l} \text{Min}[f_M - f_M^*]^T Q^{-1}[f_M - f_M^*] \\ \text{such that} \\ Cf = 0 \end{array} \right\} \quad (22.1)$$

In this problem, Q is the known $q_M \times q_M$ variance–covariance matrix for measured vector f_M . This least squares problem can be derived from different sources using the assumption that the distribution of errors follows a joint normal distribution, that is, $\epsilon_M \sim N_p(0, Q)$. Johnston and Kramer [1] presented a maximum likelihood derivation of the steady-state linear reconciliation model, and Crowe [2] showed that the same result can be derived using information theory. The above linear and steady-state data reconciliation problem was extended to the nonlinear case of estimating variables other than mass flows and to the dynamic case of estimating variables throughout time. All these extensions are discussed later and have been discussed in detail in several books [3–6]. Finally, the use of bounds (i.e., constraints on the deviations from the measurements) can also be added, especially to prevent negative reconciled values when flows are small [7,8].

Variable Classification

We recognize first that not all of the elements of vector f are measured, that some may be estimated using material

balance equations, but some may not. The latter are the ones that create problems in the solution of (22.1). Thus, the problem needs to be reformulated. We now describe a variable classification for linear systems, represented by $Cf=0$, which is needed to differentiate which set of variables can be estimated. Consider the system shown in Figure 22.1. It consists of 5 units, 3 nodes, and 11 streams. Measured streams are indicated by a small star. Consider that the holdups of all units are negligible.

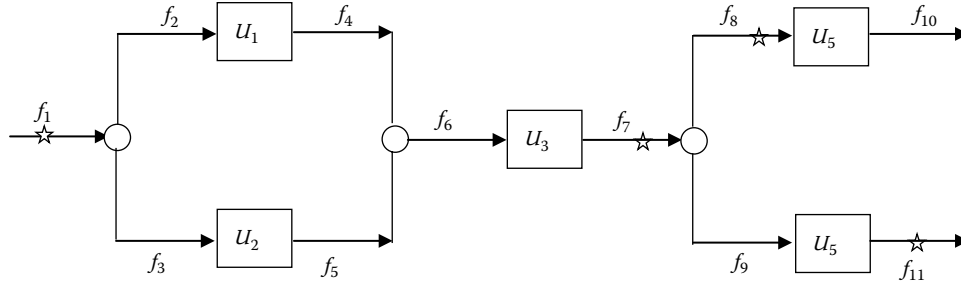
The material balances on each unit or node are as follows:

$$\left. \begin{array}{l} f_1 - f_2 - f_3 = 0 \\ f_2 - f_4 = 0 \\ f_3 - f_5 = 0 \\ f_4 + f_5 - f_6 = 0 \\ f_6 - f_7 = 0 \\ f_7 - f_8 - f_9 = 0 \\ f_8 - f_{10} = 0 \\ f_9 - f_{11} = 0 \end{array} \right\} \quad (22.2)$$

The matrix C corresponding to these balances is as follows:

$$C = \begin{bmatrix} 1 & -1 & -1 & & & & & & & & \\ & 1 & & -1 & & & & & & & \\ & & 1 & & -1 & & & & & & \\ & & & 1 & 1 & -1 & & & & & \\ & & & & 1 & -1 & & & & & \\ & & & & & 1 & -1 & & & & \\ & & & & & & 1 & -1 & -1 & & \\ & & & & & & & 1 & & -1 & \\ & & & & & & & & 1 & & -1 \end{bmatrix} \quad (22.3)$$

where the columns correspond to the variables indicated on top of the matrix and the blank spaces are zeros.

**FIG. 22.1**

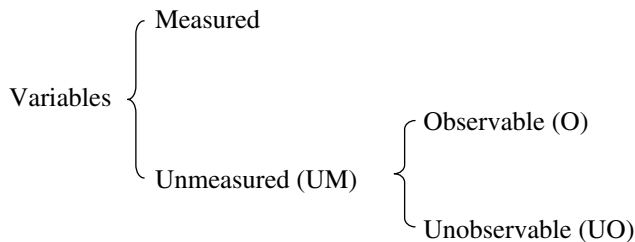
A five-unit three-node system.

For convenience, the set of state variables is divided into measured variables (f_M) and unmeasured variables (f_U). In our example,

$$f_M = \begin{bmatrix} f_1 \\ f_7 \\ f_8 \\ f_{11} \end{bmatrix}, \quad f_U = \begin{bmatrix} f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \\ f_9 \\ f_{10} \end{bmatrix} \quad (22.4)$$

The objective is now to determine which of the variables are observable. That is, for which variable is an estimate possible. Since all measured variables are estimable, this objective is really only applied to the set of unmeasured variables. In other words, since some of the unmeasured variables can be estimated using the material balance equations, one needs to determine exactly which ones are estimable.

To motivate the concepts involved, we first notice that all the flow rates after the first split (f_2, f_3, f_4 and f_5) cannot be calculated using any of the material balances. Indeed, for example, only the sum of flows f_2 and f_3 is estimable from the measurements. These variables are called unobservable variables. The rest of the unmeasured variables (f_6, f_9 , and f_{10}) can be estimated from material balances using the measured values. These variables are known as observable variables. This leads to our first classification:

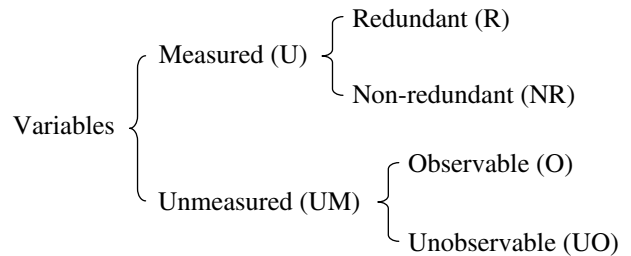


Let us turn our attention to the measured streams in Figure 22.1. If the flow rate f_1 is not measured, it can be estimated by the measurement of f_7 . However, another way is by adding f_8 and f_{11} . It is then said that the set $\{f_1, f_7, f_8, f_{11}\}$ is redundant.

Suppose now that the flow rates f_7 and f_8 are not measured. Then, because f_1 cannot be estimated by any balance equation using other measurements, we call it nonredundant. In other words, removing its measurements makes it unobservable. Thus, the following definition follows:

A measured variable is non-redundant if after removing its measurement, the variable is unobservable, i.e., it cannot be calculated using a balance equation involving the other variables of the system. Otherwise, the measured variable is called redundant.

Such redundancy is called “software or analytical redundancy.” It is a desirable property because in the case when an instrument fails, its variable can be estimated through balances. Moreover, if the number of different balances that can be used increases, there will be additional ways to calculate the variable. We therefore complete our classification of variables as follows:



Redundancy has been traditionally understood as the use of more than one instrument to measure the same variable and it is called hardware redundancy in such case. It has an effect on the accuracy and the reliability with which the estimates of the variables are obtained, but has no effect on observability of other variables.

Kalman [9] introduced the concept of observability for linear dynamic systems. Stanley and Mah [10] discussed this issue of observability of systems described by steady-state models in depth. In particular, they discuss conditions under which observability can be attained. Definitions of degree of observability and redundancy were first introduced by Maquin et al. [11,12]. Bagajewicz and Sánchez [13] unified both concepts by introducing the concept of estimability.

Canonical Representation

By using simple rearrangement of columns in matrix C , the system can be rewritten in the following way:

$$Cf = [C_U \ C_M] \begin{bmatrix} f_U \\ f_M \end{bmatrix} = 0 \quad (22.5)$$

In the case of Figure 22.1, matrices C_U and C_M are as follows:

$$C_U = \begin{bmatrix} f_2 & f_3 & f_4 & f_5 & f_6 & f_9 & f_{10} \\ -1 & -1 & & & & & \\ & 1 & -1 & & & & \\ & & 1 & -1 & & & \\ & & & 1 & 1 & -1 & \\ & & & & 1 & & \\ & & & & & -1 & \\ & & & & & & -1 \\ & & & & & & 1 \end{bmatrix}, \quad C_M = \begin{bmatrix} f_1 & f_7 & f_8 & f_{11} \\ 1 & & & \\ & -1 & & \\ & & 1 & -1 \\ & & & 1 \\ & & & & -1 \end{bmatrix} \quad (22.6)$$

The objective now is to manipulate these matrices to perform a variable classification. Specifically, in the case of measured variables, we want to distinguish redundant from nonredundant variables and in the case of the unmeasured variables, we want to distinguish observable from unobservable variables. To accomplish this, we now perform a Gauss-Jordan factorization of matrix C , concentrating on the unmeasured part only. The basic elementary steps are

1. The multiplication of two rows of matrix C by a number different from zero
2. The interchange of position of two rows (or columns) of the matrix
3. The addition of a row to another row

The resulting matrix is called the canonical form. In this canonical form, one can identify the unit matrix in the left upper corner, while the other rows, if present, are equal to zero in the same columns. This procedure was first proposed for data reconciliation by Madron [3]. For our example, this matrix is

$$C = \begin{bmatrix} f_6 & f_9 & f_{10} & f_2 & f_3 & f_4 & f_5 & f_1 & f_7 & f_8 & f_{11} \\ 1 & & & & & & & -1 & & & \\ & 1 & & & & & & & -1 & 1 & \\ & & 1 & & & & & & & & -1 \\ & & & 1 & & & & & & & \\ & & & & 1 & -1 & & & & & \\ & & & & & -1 & 1 & -1 & & & \\ & & & & & 1 & 1 & -1 & & & \\ & & & & & & & 1 & -1 & & \\ & & & & & & & & 1 & & \\ & & & & & & & & & -1 & -1 \end{bmatrix} \quad (22.7)$$

Observable Unobservable Redundant

If in addition, variable f_3 is measured, then the result is

$$C = \begin{bmatrix} f_6 & f_9 & f_{10} & f_2 & f_4 & f_5 & f_1 & f_7 & f_8 & f_{11} & f_3 \\ 1 & & & & & & & & & & \\ & 1 & & & & & & & & & \\ & & 1 & & & & & & & & \\ & & & 1 & & & & & & & \\ & & & & 1 & & & & & & \\ & & & & & 1 & & & & & \\ & & & & & & -1 & & & & \\ & & & & & & & -1 & 1 & & \\ & & & & & & & & & -1 & \\ & & & & & & & & & & 1 \\ & & & & & & & & & & 1 \\ & & & & & & & & & & -1 \\ & & & & & & 1 & -1 & & & \\ & & & & & & & 1 & -1 & -1 & \end{bmatrix} \quad (22.8)$$

Observable Redundant Non-redundant

In general, we obtain a matrix with the following structure:

$$\begin{bmatrix} \text{Unobservable} \\ \text{Observable} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & E_{UO} \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \text{Redundant} \\ \text{Non-redundant} \end{bmatrix} \begin{bmatrix} -E_{RO} & -E_{NRO} \\ -E_{RUO} & -E_{NRUO} \\ E_R & 0 \end{bmatrix} \begin{bmatrix} f_O \\ f_{UO} \\ f_R \\ f_{NR} \end{bmatrix} = 0 \quad (22.9)$$

Unmeasured Measured

where f_O , f_{UO} , f_R , and f_{NR} are the vectors of observable, unobservable, redundant, and nonredundant flows, respectively. Rewriting the system $Cf=0$ using these vectors, one obtains the following:

$$f_O = E_{RO}f_R + E_{NRO}f_{NR} \quad (22.10)$$

$$E_{UO}f_{UO} = -E_{RUO}f_R - E_{NRUO}f_{NR} \quad (22.11)$$

$$E_Rf_R = 0 \quad (22.12)$$

Equation 22.12 cannot be satisfied by the measurements and is used to adjust these measured variables through data reconciliation. Equation 22.10 allows the calculation of the observable variables. Finally, Equation 22.11 represents a system of equations that cannot be solved and involves the unobservable variables.

There are several alternatives to the above procedure. Among the most well known are the Matrix Projection [14] and the QR decomposition [15,16]. They are all equivalent.

STEADY-STATE LINEAR DATA RECONCILIATION

Consider now that only material balances are involved. In such case, only flow rates are estimated. After variable classification is performed, the data reconciliation problem shown in Equation 22.1 does not need to refer to unmeasured variables because material balance equations containing only the measured and redundant variables have been obtained. Thus, the problem can be written as follows:

$$\left. \begin{array}{l} \text{Min}[\tilde{f}_R - f_R^+]^T Q_R^{-1} [\tilde{f}_R - f_R^+] \\ \text{such that} \\ E_R \tilde{f}_R = 0 \end{array} \right\} \quad (22.13)$$

Notice that the nonredundant variables have been eliminated from consideration and a variance–covariance matrix Q_R corresponding only to redundant variables is used. The usual assumption made in industrial practice is to assume that this matrix is diagonal. However, in general, the matrix is not necessarily diagonal. In fact, there are a few methods developed to determine its components, as are discussed below.

This problem and its solution was the object of the 1965 seminal paper of this field [15], but it has older roots in statistics. It is formally a quadratic programming problem with linear equality constraints. The solution is analytical:

$$\tilde{f}_R = \left[I - Q_R E_R^T (E_R Q_R E_R^T)^{-1} E_R \right] f_R^+ \quad (22.14)$$

After reconciliation is performed, the following estimates are obtained:

$$\tilde{f}_{NR} = f_{NR}^+ \quad (22.15)$$

$$\tilde{f}_O = E_{RO} \tilde{f}_R + E_{NRO} \tilde{f}_{NR} \quad (22.16)$$

Precision of Estimates

Once the steady-state data reconciliation problem is solved, it is desired to estimate the standard error or mean square error of the estimator. The precision of the estimators is improved because their standard error decreases. Data reconciliation, via constraints on closure, reduces standard errors of estimators. Indeed, in general, if $y = \Gamma x$, then the variance of y is given by

$$\tilde{Q}_R = \Gamma Q_R \Gamma^T \quad (22.17)$$

And consequently, applying this general formula to this case, the variance–covariance matrices of the flow rate estimators are given by the following expressions:

$$\tilde{Q}_R = Q_R - Q_R E_R^T (E_R Q_R E_R^T)^{-1} E_R Q_R \quad (22.18)$$

$$\tilde{Q}_O = [E_{RO} E_{NRO}] \begin{bmatrix} \tilde{Q}_R \\ Q_{NR} \end{bmatrix} [E_{RO} E_{NRO}]^T \quad (22.19)$$

Example

Consider first the system of Figure 22.2 and the corresponding set of measured data given in Table 22.1. The reader can verify that the canonical form of this matrix is as follows:

$$C = \begin{bmatrix} & f_7 & f_2 & f_3 & f_4 & f_5 & f_1 & f_6 \\ \begin{array}{c} 1 \\ \vdots \\ 1 \end{array} & & & & & & & -1 \\ & 1 & & & 1 & -1 & & \\ & & 1 & & -1 & & & \\ & & & 1 & 1 & -1 & & \\ & & & & & 1 & & -1 \end{bmatrix} \quad (22.20)$$

From this canonical form, one can conclude that all measured streams are redundant, nonredundant variables are not present, and that S_2 , S_3 , S_4 , and S_5 are unobservable variables, whereas S_7 is observable. After data reconciliation is performed, the following results are obtained (Table 22.2).

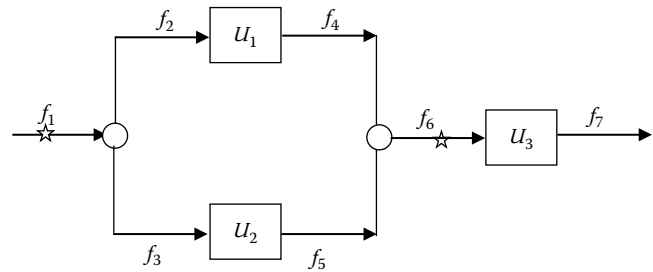


FIG. 22.2

A measurement system.

TABLE 22.1

Measurement for Figure 22.2

Stream	Measurement	True Variance
S_1	101.3	2.1
S_6	102.7	1.9

TABLE 22.2

Data Reconciliation Results for Figure 22.2

Stream	Measurement	True Variance	Reconciled/ Estimated Value	Estimated Variance of Estimator
f_1	101.3	2.1	102.07	0.9975
f_6	102.7	1.9	102.07	0.9975
f_7	—	—	102.07	0.9975

Suppose now that flow rate f_5 is also measured with a variance of 0.5. Then the new canonical form of the system matrix is as follows:

$$C = \begin{bmatrix} f_7 & f_2 & f_3 & f_4 & f_1 & f_6 & f_5 \\ 1 & & & & & -1 & \\ & 1 & & & -1 & & 1 \\ & & 1 & & & & -1 \\ & & & 1 & -1 & & 1 \\ \hline & & & & 1 & -1 & \end{bmatrix} \quad (22.21)$$

As a result of the measurement of f_5 , which is nonredundant, all the unmeasured variables become observable. Reconciled values are the same because the redundant system has not changed. The previously observable variable S_7 is not calculated using the new nonredundant measurement. Thus, its variance does not change. The rest of the variances are shown in Table 22.3.

Practical Challenges of the Steady-State Model

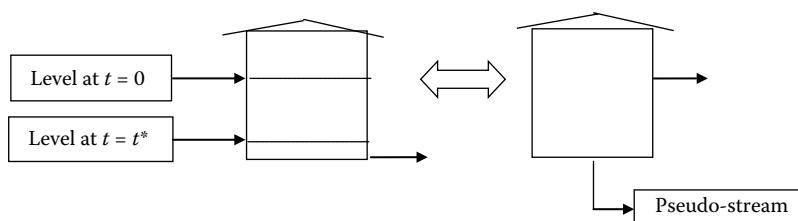
The following problems and challenges have been identified:

1. *Tank holdup measurements cannot be directly used in steady-state data reconciliation models.* In practice, to add redundancy, holdup changes are modeled as pseudo-streams. In addition, tanks are very often an important part of a chemical plant, especially in the case of refineries. It has been, therefore, of great interest for practitioners to be able to include all the transfers of raw material between tanks, from tanks to processes, or vice versa (called transactions or custody transfer when done at battery limits) as part of data reconciliation. However, tank holdup changes do not fit in the description of a steady state. Thus, when steady state is assumed over a period of time, the changes in holdup are usually divided by the time elapsed and considered as a pseudo-stream leaving (or entering) the system (Figure 22.3). If the level is measured, then the pseudo-stream is considered a

TABLE 22.3

Data Reconciliation Results for Figure 22.2

Stream	Measurement	True Variance	Reconciled/ Estimated Value	Estimated Variance of Estimator
f_1	101.3	2.1	102.07	0.9975
f_6	102.7	1.9	102.07	0.9975
f_7	—	—	102.07	0.9975
f_5	33.8	0.3	33.8	0.3
f_3	—	—	33.8	0.3
f_2	—	—	68.27	1.2975
f_4	—	—	68.27	1.2975

**FIG. 22.3**

Use of pseudo-streams in tanks.

measured stream. Real streams entering or leaving the tank are considered separately. Because transactions between tanks are reported, this conversion has been very useful in adding substantial amount of redundant streams to refinery installations and has proven to be valuable to perform refinery-wide oil accounting and oil-loss assessment.

2. *Plants are never truly at a steady state.* As practitioners have resorted to perform averages of several measurements to obtain one single number per stream to use in the objective function of (22.13), they are averaging process variations as well. Process variations, due to all kinds of disturbances, are often ignored and lumped into one so-called “measurement” for each instrument. The assumption that gives some validity to such a procedure is that the plant is at this so-called pseudo-steady state, where, loosely speaking, state variables fluctuate and/or drift, “but not that much.” From the purely theoretical point of view, this practice has been considered questionable. In principle, accumulation terms are not considered in these models. They are only included in integral form using the idea of considering accumulation as an input or output stream of a unit, a procedure that cannot capture the changes due to the dynamics of the system. An average, no matter how weighted or transformed, cannot possibly be validated as a “pseudo-steady-state” value, the criticism goes.

In spite of all these criticisms, steady-state data reconciliation often produces results that are acceptable for the practitioner, who does not seem to see the averaging of slowly drifting systems as a problem.

However, users of data reconciliation software wonder how to most effectively deal with this averaging: Should one average in shorter periods of time? How small should be the time interval? and so on. Practitioners are also increasingly aware of this deficiency, and it will probably become an issue when pressure mounts to get more accurate and realistic figures for accounting, control, and monitoring purposes. This issue is later covered in more detailed when the detection of biases and leaks is discussed.

The answer to this could be to apply dynamic data reconciliation methods, which are briefly reviewed later. However, quite recently, a connection between dynamic and steady-state data reconciliation was found [18] for the case of linear systems (material balances only). Specifically, it is proven that, provided the variance–covariance matrix is diagonal, then *the results of the steady-state data reconciliation of averaged values are equal to the average of the dynamic data reconciliation values* (which are obtained for all time intervals). The aforementioned article also explores the effect of holdup and concludes that the steady state still constitutes a

reasonable approximation. In a study by Bagajewicz and Gonzales [19], another dynamic data reconciliation model [20] was applied to nonsmooth signals, proving that the approximation is also good in these cases. The above findings provide a reasonable argument for continuing using the technology available in commercial software. However, when process variations are too large, certain averaging practices can render larger variances upon reconciliation than those of the measurements themselves [21]. In such circumstances, it is preferred not to reconcile.

One practical approach has also been to perform data reconciliation using steady-state models only after the system has been identified to be at steady state. For this, various methods have been proposed [22–27]. The issue is discussed in detail in a recent book by Bagajewicz [28].

For processes with high sampling rate for all variables, one can collect data over a short duration before the process significantly deviates from some steady state. However, for processes with slow sampling rates relative to process changes, it will not likely be possible to collect multiple samples at a single steady state. What one wants to do is to take enough samples so that process variations are somehow averaged. Consider the following measurement model for pseudo-steady state:

$$f_R^+ = f_R + \lambda_R + \varepsilon_R + \delta_R \quad (22.22)$$

where

f_R is the “steady-state” value

λ_R is the process variation around f_R

ε_R is the random error

δ_R is the bias

This model is valid for each instance of time a measurement is made. Although λ_R is assumed to change independently over time, it is likely to be serially correlated in real processes (i.e., changing according to some pattern). This assumption will not likely be too critical as long as the mean of λ_R is close to zero. One can ensure this by allowing enough time for the process to cycle around f_R with about the same magnitude of positive deviations as negative deviations. More guidelines regarding the sample size are discussed below in connection to gross error detection.

3. *It has been a problem to pick the values of the variance–covariance matrix Q .* Typically covariances are ignored and a diagonal form of Q is used. In the absence of hard variance data, as questionable as this practice may seem, these variances are chosen in industrial practice using vendor information or in some cases, the standard deviation of the measurements.

If measurements are independent then they are not correlated (which is nonetheless not likely in the case of plant data). The variance–covariance matrix Q is diagonal and the elements of the diagonal are the variances of the individual measurements. In the case of steady-state data reconciliation, estimates of these values can be obtained by calculating the variance of the distribution of data around the mean value. This is called the *direct method* [29]. This is a valid procedure if the two aforementioned assumptions, independence and steady state, hold. In addition, outliers must not be present or they should be removed.

Since the system is never truly at steady state, the above formulas of the direct method incorporate the process variations, that is, the variance of the natural process oscillations or changes as part of the measurement variance. In a simple case like a ramp function, for example, the variance will be a composite of one half the changes in the true value of measured value during the sampling interval and the true variance. To ameliorate this problem and to assess the existence of variable interdependence (nondiagonal variance matrix), the indirect method was proposed [29]. In this method, the sum of the squares of the off-diagonal elements of the variance–covariance matrix of the constraint residuals ($r = E_R f^+$) is minimized. The method was later slightly modified using an iterative procedure based on the solution of a nonlinear optimization resulting from using a maximum likelihood estimator [30]. Finally, Keller et al. [31] extended this work to nondiagonal covariance matrices. All these approaches still suffer from the problem that they do not consider the possible presence of outliers. Chen et al. [32] proposed the application of an M-estimator that applies a weight to each data based on its distance to the mean. They called this the *robust indirect method*. The discussion of the estimation using dynamic data is omitted.

While all these methods for variance estimation have been tested using computer simulations and have shown their power in these controlled experiments, there is no assessment of how they behave in practice. In particular, data reconciliation software notoriously lacks of any module to perform such estimation, and there are no published results regarding the efficiency of these methods in practice.

4. *There is confusion in practice as of what boundaries to choose for reconciliation.* A large complex plant, especially refineries consists of several different processing units and a set of tanks to store raw materials and products. In addition, sometimes, the whole company has different sites that are interconnected through transportation systems (pipelines, trucks, etc.). As different units may have different levels of redundancy, the question arises whether one should take the whole

complex and perform a large-scale data reconciliation, risking that gross errors will smear good data, or perform reconciliations at the level of each unit first. A strategy to deal with this problem is detailed in a book by Bagajewicz [28].

BIAS AND LEAK DETECTION

The sources of gross errors are instrument biases, leaks, and true outliers (occasional measurements that depart significantly from all other measurements). Departures from steady state have been included in this list in the past. As discussed above, some of these have been clarified for linear systems [13]. The challenging task is to

- Identify the existence of gross errors (biases, leaks, and outliers)
- Identify the gross errors location
- Identify the gross error type
- Determine the size of the gross error and eliminate its influence on the final estimates

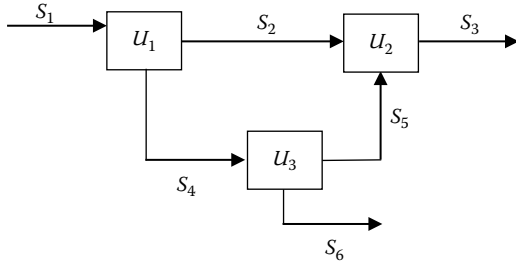
There exists at least one method that allows the detection of the existence of gross errors. Some of the methods for gross error identification are to a certain extent capable of discerning the location and type. After the gross errors are identified, two responses are possible and/or desired:

- Eliminate the measurement containing the gross error, or
- Correct the measurements or the model and run the reconciliation again

The first alternative is the one implemented in commercial software that in general only considers biases and incapable of detecting leaks. This leaves the system with a smaller degree of redundancy and, as we saw, the quality of the reconciliation is deteriorated. If one is able to identify the gross errors and obtain a reliable estimate of their values, then the second alternative becomes appealing because redundancy is not lost. Some methods perform such task with reasonable success.

Identification Using Closure

One popular way of identifying measured variables with significant bias and nodes with significant leaks is testing for material and energy balance closure in nodes (interconnecting point or units). When closure is achieved, the streams entering and leaving the node are concluded to be unbiased and leaks at this node are concluded to be negligible. This strategy is illustrated using the flowsheet of Figure 22.4, which is assumed to be in a steady state and where, for simplicity, random errors are zero.

**FIG. 22.4**

Process network used to illustrate a closure strategy.

Consider that there are no leaks and that measurement bias exists only in the measured mass flow rate for Stream 2. Then the measurements f^+ are given by

$$\left. \begin{aligned} f_i^+ &= f_i & \forall i \neq 2 \\ f_2^+ &= f_2 + \delta_2 \end{aligned} \right\} \quad (22.23)$$

where

f_i are the true values
 δ_i are the biases

Overall mass balances on the nodes give

$$\left. \begin{aligned} r_1 &= f_1^+ - f_2^+ - f_4^+ \\ r_2 &= f_2^+ - f_3^+ + f_5^+ \\ r_3 &= f_4^+ - f_5^+ - f_6^+ \end{aligned} \right\} \quad (22.24)$$

Combining (22.23) and (22.24), and recognizing that at steady state, the true flow rate values provide node closure when mass flows are used, one gets

$$\left. \begin{aligned} r_1 &= f_1 - f_2 - \delta_2 - f_4 = -\delta_2 \\ r_2 &= f_2 + \delta_2 - f_3 + f_5 = \delta_2 \\ r_3 &= f_4 - f_5 - f_6 = 0 \end{aligned} \right\} \quad (22.25)$$

If closure is concluded on testing, then all measured variables (i.e., stream flow rates) entering and leaving the node or combination of nodes are concluded to be unbiased (i.e., $\delta=0$) and the nodes are concluded not to have significant leaks. Error cancellation, that is, cases where biases or leaks exist but add to zero, is typically not likely and is assumed to not occur [21]. Nevertheless, in the strict sense, this procedure only makes conclusions regarding zero biases and leaks, but it does not make conclusions regarding them being nonzero.

Selection of Nodes

For the process in Figure 22.4, there are seven balances one could test for closure. They are balances around the individual

units, U_1 , U_2 , and U_3 and balances around combination of nodes, that is, U_1U_2 , U_1U_3 , U_2U_3 , and $U_1U_2U_3$. Assume that one decides to test closure for the individual units U_1 , U_2 , and U_3 , then the nodal combination of $U_1U_2U_3$. With measurement random errors removed, the conclusion can be known for each test and there is no need to rely on statistical hypothesis testing. Thus, (22.27) gives the results for the balances around the individual nodes and $r_{123} = r_1 + r_2 + r_3 = 0$. Therefore, no conclusions are made regarding closure for nodes U_1 and U_2 . In contrast, since $r_3 = 0$, measured streams S_4 , S_5 , and S_6 are concluded to be unbiased (i.e., $\delta_4 = \delta_5 = \delta_6 = 0$). Next, the test closure from r_{123} concludes that $\delta_1 = \delta_3 = \delta_6 = 0$. Thus, from this strategy of testing, the overall conclusion is that $\delta_1 = \delta_3 = \delta_4 = \delta_5 = \delta_6 = 0$. As stated above, it is not proper for nodal testing methods to conclude that $\delta_2 \neq 0$. One could just simply state that it may not be zero. Therefore, a statistical test is needed to determine if such an overall conclusion can be made in the presence of measurement error. In order for this approach to have practical merit, tests and strategies are needed to control error rates for false conclusions. This discussion is restricted for simplicity to the case of pseudo steady state.

Hypothesis testing has been used to detect and identify gross errors. We now discuss a few of these methods. We also assume that averages of a certain number of measurements (N_M) have been taken. Thus, an average of all residual vectors (denoted by r) is also defined.

Global Test

The null hypothesis (H_0) is that no bias is present. Thus, if r is the vector or residuals of the expected value of r ($= E_R f^+$) is zero, that is, $E[r] = 0$ and N_M is the number of measurements, then, in the absence of gross errors, and under the assumption of multivariate normality for f^+ , the following statistics

$$\chi_m^2 = N_M r^T J^{-1} r \quad (22.26)$$

follows a χ^2 distribution with m degrees of freedom, where m is the number of rows of E_R . Matrix J is the variance-covariance matrix of r and is given by

$$J = \text{Cov}(r) = E_R Q_R E_R^T \quad (22.27)$$

where Q_R is the variance-covariance matrix of f^+ .

The test therefore consists of rejecting H_0 if $N_M r^T J^{-1} r$ is larger than a threshold value $\chi_{m,\alpha}^2$, the upper α th percentile of the χ^2 distribution with m degrees of freedom.

Notice that this test can be performed without actually implementing data reconciliation. If this number falls within the nonrejection region, then the null hypothesis is not rejected, that is, no gross error is suspected. On the other hand, if $N_M r^T J^{-1} r$ is larger than the critical value, a gross

error cannot be ruled out, although in practice it is usually assumed it has been detected. Its location, however, cannot be determined using this test.

Nodal Tests

In the absence of gross errors, the constraint residuals r follow an m -variate normal distribution (m is the rank of E_R), with zero mean. Therefore, the following test statistic

$$Z_i^N = N_M^{1/2} \frac{r_i}{\sqrt{J_{ii}}} \quad (22.28)$$

follows a standard normal distribution, $N(0,1)$, under H_0 and $Z_{\alpha/2}$ is the $100(1-\alpha/2)$ th percentile of the standard normal distribution.

Therefore, if Z_i^N is larger than the critical value ($Z_{\alpha/2}$) based on the level of significance α , then one concludes that there is at least one gross error in the set of measurement that participates in the corresponding nodal balance. Notice that $\alpha/2$ is used instead of α because the test is a two-tail test, that is, it tests for large positive or negative departures from H_0 .

Assume that there are k tests that one makes on a single set of data. Multiple tests are then being made using the same critical value, and for this reason, the likelihood of a type I error (wrongly reject H_0 , that is, declare the existence of a gross error when there is none) increases. Mah and Tamhane [33] proposed the use of a new smaller level of significance (β) derived using the Sidák inequality [34]. This new level of significance is given by

$$\beta = 1 - (1 - \alpha)^{1/k} \quad (22.29)$$

that is, to use $Z_{\beta/2}$ instead of $Z_{\alpha/2}$ as a threshold value. The Bonferroni adjustment for multiple testing is

$$\beta = \frac{\alpha}{k} \quad (22.30)$$

which is another alternative [35]. Note that when k is large, (22.29) approaches (22.30), but even for small values of k they are very similar. Since Sidák's inequality assumes independent tests and Bonferroni does not, the use of the Bonferroni correction is recommended.

Since the square of a variable that follows a standard normal distribution follows a χ^2 one, then the statistics

$$Z_i^C = N_M \frac{r_i^2}{J_{ii}} \quad (22.31)$$

can also be used to test the presence of gross errors. Thus, H_0 is rejected if $Z_i^C \geq \chi_{m,\alpha}^2$.

Tests Using Linear Combination of Nodes

It was discussed that different combinations of nodes lead to different conclusions that one can make. A particular combination of nodal residuals can be represented by $l_k^T r$. For example, for the network of Figure 22.4, the nodal residual combination $r_1 + r_2$ is obtained using $l_k^T = (110)$. There are $2^m - 1$ of such vectors. Thus, the nodal test statistic based on the standard normal distribution for these combinations becomes

$$Z_k^{NI} = N_M^{1/2} \frac{|l_k^T r|}{\sqrt{l_k^T J l_k}} \quad (22.32)$$

while the nodal test statistic based on the χ^2 distribution becomes

$$Z_k^{CI} = N_M \frac{(l_k^T r)^2}{l_k^T J l_k} \quad (22.33)$$

In principle, one could perform this test on all combination of nodes, but since the number of such tests can be very high, Rollins et al. [36] proposed guidelines to disregard certain nodes in an effort to maximize power. Tests that are suitable for the case where variances and covariances are not known were also developed [37].

Note that the test based on the standard normal distribution (Equation 22.32) is independent of the number on tests contrasting to the one based on the χ^2 distribution (Equation 22.33), which depends on the number of tests. When the number of tests (k) is small, the nodal test based on the normal distribution will be more powerful than the one based on the χ^2 distribution. In other words, the Bonferroni test will be more powerful if $(Z_{\alpha/2} k)^2$ is less than $\chi_{m,\alpha}^2$. For $k \leq m$, the test based on the normal distribution is usually more powerful than the χ^2 test. However, when $k > m$, the test based on the χ^2 distribution can be more powerful. This is true in particular when $k \gg m$.

Finally, Crowe [38] proposed a linear transformation for which the nodal test has maximum power, that is, minimum of failures to identify existing gross errors, which are known as type II errors. Such test is equivalent to the nodal test given by Equation 22.32.

Measurement Test

The measurement test (MT) is based on the vector of measurement adjustments (or corrections) denoted by Δf_R and is defined by

$$\Delta f_R = f_R^+ - \tilde{f}_R \quad (22.34)$$

The test is based on the assumption that the random errors for measurements are independently and normally distributed.

Under the null hypothesis (H_0), the expected value of Δf_R is zero. That is, the following statistic

$$Z_i^{MT} = \frac{\Delta f_{R,i}}{\sqrt{(\tilde{Q}_R)_{ii}}} \quad (22.35)$$

follows a normal distribution with zero mean under H_0 .

Since multiple tests are involved, the test should be based on a level of significance given by β obtained from Equations 22.29 or 22.30.

Finally, Mah and Tamhane [33] also proposed a linear transformation for which the MT has maximum power under some rather restrictive assumptions. The linear transformation consists of using the new transformed adjustments as follows:

$$d_R = Q_R^{-1} \Delta f_R \quad (22.36)$$

and therefore the maximum power test statistics, which is used the same way as the regular MT, is given by

$$Z_i^{MMP} = \frac{|d_{R,i}|}{\sqrt{(\tilde{Q}_R)_{ii}}} \quad (22.37)$$

The MT is “statistically inadmissible,” and therefore unacceptable, because it has a property called “inconsistency.” This means that it does not give the correct answer with infinite sampling. In short, with the measurement error removed (i.e., using the deterministic solution) and one biased variable, the test may point to the wrong variable, which is incorrect. In order for a method to be admissible, it has to give the correct deterministic solution. Nevertheless, the method has been included here because of its widespread use in the field for the identification of biases.

Generalized Likelihood Ratio Test

The alternative hypothesis in this test consists of a particular bias in a stream associated to a node or a leak [39], that is, H_1 : $\mu_r = b h_i$, where μ_r is the expected value of the node residual, g_i a vector in the direction of a bias ($h_i = E_R e_i$), or in the direction of a leak ($h_i = m_i$), and b is the size of this gross error. This is an uncommon way of posing an alternative hypothesis because it contains an unknown number (b). Nevertheless, this is equivalent to H_0 : $\mu_r = 0$ and H_1 : $\mu_r \neq 0$. The test is based on finding the supremum of the likelihoods of each hypothesis, that is,

$$\lambda = \sup \frac{\Pr\{r|H_1\}}{\Pr\{r|H_0\}} \quad (22.38)$$

where the supremum is computed over all possible bias and leaks. When the probability distributions used are normal, the test consists of computing

$$T = 2 \ln \lambda = \sup_{\forall i} \frac{(h_i^T Q_R^{-1} r)^2}{h_i^T Q_R^{-1} h_i} \quad (22.39)$$

which is compared with the corresponding threshold.

Unlike the MT, this test appears to be consistent when one gross error is present. However, when multiple gross errors are present, the test procedure can fail under deterministic conditions, and is therefore not statistically consistent.

Principal Component Test

Tong and Crowe [40] proposed the use of principal components, that is, Eigenvalue decomposition of the variance of the residuals ($E_R Q_R E_R^T$), given by

$$\Lambda_r = U_r^T (E_R Q_R E_R^T) U_r \quad (22.40)$$

where U_r is the matrix of orthonormalized eigenvectors of $E_R Q_R E_R^T$. Accordingly, the principal components are

$$p_r = (\Lambda_r^{-1/2} U_r)^T r \quad (22.41)$$

which are tested against the normal distribution. Once a suspect has been identified, the error is traced back to the corresponding node by looking at the largest contributions to this component. A similar MT using principal components was developed [28].

Sample Size

Issues related to process variations were discussed above. However, no matter which strategy one chooses for data collection, it is important to recognize that sampling should be limited. If the number of samples N_M is too large, the power (i.e., probability of detecting a nonzero bias δ or leak λ) will be too large [29]. Since all measured streams are likely to have some degree of bias, very high power will declare that all streams bias, which will not be a helpful conclusion. Thus, one should attempt to control power by selecting the sample size a priori based on control of the Type II error (i.e., false detection) level. Thus, too much power can be a problem—an issue discussed by Chen et al. [41].

MULTIPLE GROSS ERROR IDENTIFICATION

In the case where multiple gross errors (biases, leaks and outliers) are present, a strategy to identify them is needed. We briefly discuss two strategies, the first one of which is widely used in commercial software.

Serial Elimination

In this strategy, a selected test (MT or principal component analysis) is coupled with an elimination strategy. If the test flags the existence of gross errors, then a strategy is proposed to identify one or more variables, which are the “most suspected ones.” The measurements of these variables are eliminated and the test is run again, even if that implies performing the data reconciliation again. Commercial software has used the MT in this strategy with relatively good success, but with some troubling counter examples. In the presence of multiple gross errors, the MT usually (but not always) gives the largest value for the variable where the gross error exists and this variable is singled out for elimination. The reader is reminded of the limitations regarding the MT stated above.

Unbiased Estimation (UBET)

This method renders unbiased estimators [21]. Candidate gross errors are identified and a simple formula is used to estimate their size in order to obtain unbiased estimates for process variables and leaks.

UBET is developed from the balance residuals r and its expected value

$$\mu_r = A\delta + M\gamma \quad (22.42)$$

where γ represents the possible leaks. By partitioning A , M , δ , and γ , one gets

$$\mu_r = \begin{bmatrix} A_{11} & 0 \\ A_{21} & M_{22} \end{bmatrix} \begin{bmatrix} \delta_1 \\ \gamma_2 \end{bmatrix} = C_1 \theta_1 \quad (22.43)$$

Finally, by introducing the vector l_i by means of $l_i^T = e_i^T C_1^{-1}$, one obtains

$$l_i^T \mu_r = e_i^T \theta_1 \quad (22.44)$$

Thus, $l_i^T r$ (for all i) are unbiased estimators of the components of δ and γ contained in θ_1 . Originally, the method was proposed to be preceded by an identification step, at which point the procedure would stop. Some difficulties of this method have been studied, and measures to overcome them have been suggested [30]. They are connected to the issue of equivalency of gross errors, which is discussed below. In particular, Bagajewicz and Jiang [44] showed that UBET can be used by assuming m locations that correspond to a spanning tree of the graph of the flowsheet. Once the errors are estimated, hypothesis testing can proceed. These errors are then equivalent to several alternative sets.

Uncertainties

We now illustrate some uncertainties that arise in practice, especially when the MT is used. We reiterate the limitations

TABLE 22.4

Measured Flow Rates and Variances

Stream	Measured Flow Rate	True Variance
f_1	10.03	0.1
f_8	5.99	0.03
f_{11}	3.99	0.16

of this test made above, but we include this discussion in view of its widespread use in commercial software.

In many cases (actually very often), the MT gives the same value for many variables. Consider the system of Figure 22.1 and assume that only S_1 , S_8 , and S_{11} are measured. They are redundant. Assume the values given by Table 22.4.

If one adds a bias of size 5 to the measurement of the flow rate f_1 , making the measurement 15.03, the reconciliation renders

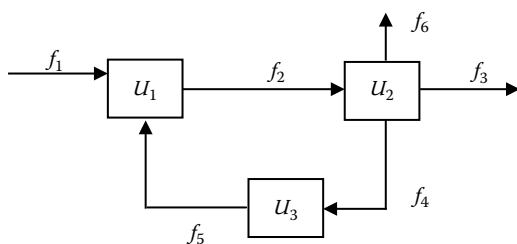
$$\tilde{f}_R = \begin{bmatrix} 13.29 \\ 6.51 \\ 6.78 \end{bmatrix} \quad (22.45)$$

The values of the Z-statistics for the MT are all the same. These values are $Z=9.37$. This illustrates the uncertainty of which variable should be eliminated in the case where the strategy is based on the elimination of the variable with the largest Z-statistics. This poses a problem for the serial elimination strategy because there is no criterion to determine which one of these three measurements should be eliminated. This phenomenon was first pointed out by Iordache et al. [42] and is now well understood [43]. In the case of two gross errors, it happens because the corresponding columns of E_R are proportional. Indeed, the last two columns (corresponding to S_8 and S_{11}) are equal, and the proportionality constant between the first and the two last columns is (-1) . Cases involving more streams are explained by Bagajewicz [28]. As it will be shown, if one understands the nature of these uncertainties, the serial elimination strategies are not flawed (in this sense) after all.

Equivalency of Gross Errors

Two sets of multiple biases are equivalent when they have the same effect in data reconciliation, that is, when eliminating either one leads to the same value of the objective function. Therefore, the equivalent sets of gross errors are theoretically indistinguishable. In other words, when a set of gross errors is identified, there exists an equal possibility that the true locations of gross errors are in one of its equivalent sets.

Consider the case of Figure 22.5 and the measurements given in Table 22.5. Random errors have been eliminated in this table so that the phenomenon is clearly visible.

**FIG. 22.5**

Flowsheet to illustrate gross error equivalences.

TABLE 22.5

Illustration of Equivalent Sets in $\{f_2, f_4, f_5\}$ of Figure 22.5

		f_1	f_2	f_3	f_4	f_5	f_6
Measurement		12	18	10	4	7	2
Case 1	Reconciled data	12	18	10	6	6	2
	(Bias in f_4, f_5)				-2	1	
Case 2	Reconciled data	12	19	10	7	7	2
	(Bias in f_2, f_4)				-1	-3	
Case 3	Reconciled data	12	16	10	4	4	2
	(Bias in f_2, f_5)				2	3	

We now analyze the set $\{f_2, f_4, f_5\}$. As it is shown in Table 22.5, three different sets of biases and reconciled values can explain these measurements (Cases 1, 2, and 3). Therefore, without a priori additional knowledge as to where the biases might be, or some additional information about the reconciled values, the three cases are impossible to be correctly identified.

We noticed that every case has two biases, which is not a coincidence. These are called *equivalent sets*. Every equivalent set has a certain *minimum* number of biases that can represent all situations. Indeed, there are infinite numbers of sets of three gross errors that can be represented by two biases. Equivalent sets can be identified using the method developed by Bagajewicz et al. [44]. Several intricacies of the gross error equivalency were discussed in this chapter. In addition, in the presence of random errors, one can define quasi-equivalency as an approximation to true equivalency to detect leaks by searching for an equivalent set of biases [43].

This has very important consequences in practice. For example, if the biases are in reality in $\{f_2, f_4\}$, any strategy for gross error detection can identify *with equal probability* any of the three cases. The impact of this is not only in the location and size of the biases but also in the reconciled values. Its importance in production accounting is paramount.

Once all the equivalent sets have been identified, one may want to determine their sizes of the gross errors in one equivalent set, when the sizes in another equivalent set have been calculated with some gross error detection and estimation method. In addition, one wants to know what the new values of the reconciled streams are.

Thus, the existence equivalency of gross errors leads to several unwanted situations. One of these situations is that the gross error detection methods will identify the wrong set of errors leading to the removal of good measurements and therefore incorrectly leave biased measurements that will smear the reconciliation. The issue and some remedies (error size estimation, a data base of failures, addition of nonlinear equations, change in instrumentation maintenance schedules, and instrumentation upgrade) are discussed by Bagajewicz [28].

Gross Error Size Estimation

Gross errors can be estimated using several methods. The oldest method is the compensation model. It consists of assuming a set of instrument biases and introducing the new “measurement”

$$f_R^+ = \tilde{f}_R + \delta \quad (22.46)$$

where δ is a vector that contains biases for bias candidates in a limited number of positions. The rest of the elements of δ that are not candidates are zero. Conducting a data reconciliation using these measurements renders estimates, which in turn can be used to obtain the values of the nonzero elements of δ . In order to avoid running into singularities, one equivalent set needs to be picked for the nonzero elements of δ . The complete procedure is illustrated by Bagajewicz et al. [44].

Jiang and Bagajewicz [45] used this model in a serial strategy where the MT is used at each step to identify a candidate gross error, which is added to a list of candidates. In contrast with other serial procedures where the sizes of the gross errors identified are determined at each step, this procedure evaluates the size of all the gross errors in the candidate list at each step. Finally, the UBET method presented earlier provides the size of gross errors.

Use of Alternative Objective Functions

Some alternative objective functions that are capable of handling the gross errors in the data simultaneously with data reconciliation have been proposed. In this chapter, we cite only those that have reached publicly offered software. Tjoa and Biegler [46] proposed a mixture distribution as the objective (likelihood) function:

$$\text{Min} \left\{ aP(\epsilon, \alpha) + (1-a)P\left(\epsilon, \frac{\sigma}{b}\right) \right\} \quad (22.47)$$

where

a is the probability associated with the absence of gross errors

σ^2 is the variance of such errors

$b^2\sigma^2$ is the variance of the biases

A test follows the reconciliation to determine the gross error presence. The apparent shortcomings of this approach [47] were addressed by Albuquerque and Biegler [48], who proposed the use of the Fair function:

$$\rho(\epsilon) = c^2 \left[\frac{|\epsilon|}{c} - \log \left(1 + \frac{|\epsilon|}{c} \right) \right] \quad (22.48)$$

where c is a parameter, and it has the advantage of being convex with continuous first and second derivatives.

In turn, Johnston and Kramer [1] proposed using the Lorentzian distribution:

$$\rho(\epsilon) = \frac{1}{1 + \epsilon^2/2} \quad (22.49)$$

which is as a robust estimator because it has the ability to filter large gross errors. Alternative robust data reconciliation schemes have been proposed [49–54]. All these methods are centered on outlier detection due to a large variance and not on bias detection. Finally, Bayesian methods were utilized to perform data rectification, a technique that is close to data reconciliation [55,56].

Accuracy in the Context of Data Reconciliation

Accuracy of an instrument is defined as the sum of the systematic error plus the precision of the instrument [57]:

$$a_i = \delta_i + \sigma_i \quad (22.50)$$

where a_i , δ_i , and σ_i are the accuracy, systematic error, and precision (square root of variance) of the mean of a certain number of repeated measurements made by a meter on variable i . The accuracy and precision are only equivalent in the absence of systematic errors or biases. When data reconciliation is used, any gross error in the system generates induced biases in all other estimators of redundant and observable variables.

We now concentrate on defining accuracy using a new concept: the induced bias. Let the measurement vector be described by

$$f_R^+ = f_R + \epsilon_R + \delta_R \quad (22.51)$$

which is similar to (22.22) but without the process variation λ_R . Now, using the common assumption that $\epsilon_R \sim N_p(0, Q_R)$ and because $E_R f_R^+ = 0$, the maximum likelihood estimate of f_R is given by

$$f_R^* = f_R - Q_R E_R^T (E_R Q_R E_R^T)^{-1} E_R f_R^+ \quad (22.52a)$$

with expected value given by

$$E[f_R^*] = f_R + \delta_R - Q_R E_R^T (E_R Q_R E_R^T)^{-1} E_R \delta_R \quad (22.52b)$$

Bagajewicz [58] defined induced bias as the one that is observed in all reconciled values due to undetected biases. Thus, from (22.52), the vector of induced biases is given by

$$\hat{\delta}_R = \left[I - Q_R E_R^T (E_R Q_R E_R^T)^{-1} E_R \right] \delta_R \quad (22.53)$$

Any bias detection technique is only effective to detect gross errors of a size that is above a certain threshold. Below such threshold, biases are not detected and they remain as induced biases in the reported estimators. Thus, small induced biases will always persist. This leads to a new definition of accuracy:

$$\hat{a}_i = \hat{\sigma}_i + \delta_i^* \quad (22.54)$$

where $\hat{a}_i$, δ_i^* , and $\hat{\sigma}_i$ are the accuracy, the maximum undetected induced bias, and the precision (square root of variance $\hat{Q}_{ii}$) of the estimator, respectively.

Bagajewicz [58] offered a way of computing this accuracy. However, this value has been found to be too conservative as a bias is not always at its maximum value. Rather, biases have a distribution (when they are random), or they grow in time following some pattern, so it is likely that the value that needs to be used in the definition is an expected undetected induced bias ($E[\hat{\delta}_i]$) and not the maximum possible undetected bias δ_i^* . Such stochastic accuracy definition was introduced by Bagajewicz and Nguyen [59], who used Monte Carlo simulations to determine its value. Comparisons between both approaches and discussions regarding the effect of gross error equivalency are included in the book by Bagajewicz [28].

Nonlinear Steady-State Data Reconciliation

Nonlinear steady-state data reconciliation is represented by

$$\left. \begin{array}{l} \text{Min}[\tilde{x}_M - z_M]^T Q^{-1} [\tilde{x}_M - z_M] \\ \text{such that} \\ g(\tilde{x}) = 0 \end{array} \right\} \quad (22.55)$$

For process plants, z_M includes the typical state variables, flow rates, concentrations, temperatures, and pressures, and the model $g(\tilde{x})$ can include any type of unit operations and equipment. In addition, $\tilde{x}$ usually contains parameters that are not measured directly and often requires estimation from process data.

Several methods have been proposed to solve this problem, especially when $g(\tilde{x})$ is bilinear. Since the solution is supposed to be close to the measurement (unless gross errors are present), then one can linearize $g(\tilde{x})$, perform a classification of variables, and extract the redundant system of equations. Once this is done, one can solve successively updating the Jacobian in each iteration until convergence is achieved. Other approaches using nonlinear programming, such as the popular Sequential Quadratic Programming codes can also be used.

Commercially available software performs nonlinear steady-state data reconciliation to a good extent. Some have sophisticated property estimation routines such that they are capable of performing material, energy, and component balances, all of them simultaneously; reconciling temperature, pressure, concentration, and flow rates; and performing parameter estimation, for example, the heat transfer coefficient of heat exchangers. In the absence of systematic errors and leaks in the system, there is no reason why these models cannot be made as sophisticated and complex as the optimization techniques used to solve them can handle. However, software vendors are reluctant to introduce such models. One of the reasons is certain conviction that undetected gross errors may still be largely amplified by the reconciliation, especially when nonlinearities correspond to very nonideal systems. If one linearizes the system ($g(\tilde{x})$) around the design or operating point and ignores terms of higher order, one can obtain a linear problem. This approach has been proposed by several authors through time [15,60–64]. The issue is discussed and illustrated by Bagajewicz [28] for a chemical reactor system.

DYNAMIC DATA RECONCILIATION

Early work in dynamic data reconciliation is rooted in the problem of process state estimation using the concept of filtering. Recently, the problem has been solved using the concept of model-based data smoothing. Consider the three types of state estimation problems that are illustrated in Figure 22.6. Assume an estimation of the state of the system is desired at time t .

When only measured values prior to the time of prediction t are used, including the measurement at time t , the estimation

is called filtering. When time t is not included, the estimation is called prediction, and finally, when data for times larger than t are used, the estimation process is called smoothing. Finally, when discrete measurements are used serially over time, the estimators are called discrete estimators.

Consider now a discrete system whose behavior at the k th time instant is given by the following model:

$$x_k = F_{k-1}x_{k-1} + w_k \quad (22.56)$$

where

x_k is a vector of system states

w_k is called process noise and has zero mean and variance R_k

For example, the discrete model for a constant is $x_k = x_{k-1}$ for all k . Consider also the following measurement model:

$$z_k = H_k x_k + v_k \quad (22.57)$$

where

v_k is a zero mean random error with variance S_k

z_k is the vector of measurements at time t_k , namely $z_k = [z_{k,1}, z_{k,2}, \dots, z_{k,m}]^T$

A linear estimator is constructed as follows:

- The a priori estimate at time t_k is obtained using the system equation:

$$\hat{x}_k^{(-)} = F_{k-1} \hat{x}_{k-1}^{(+)} \quad (22.58)$$

where we have ignored the process noise.

- Given the estimate at the k th time instant denoted as $\hat{x}_k^{(-)}$, we seek an update estimate $\hat{x}_k^{(+)}$ based on z_k using the following linear, recursive form:

$$\hat{x}_k^{(+)} = K_k^{(-)} \hat{x}_k^{(-)} + K_k^{(z)} z_k \quad (22.59)$$

One such model is the well-known Kalman filter [9], which in addition to providing the estimates provides an estimate of their variance–covariance matrix. If one assumes that the system is at steady state, then the model is $x_k = x_{k-1}$, a simplified Kalman filter can be obtained. If in addition, one also assumes that the model satisfies the balance equations ($Cx_k = 0$), a quasi-steady-state Kalman filter is obtained [65]. Other variants of this model have been proposed. For example, one could use neural networks in conjunction with the extended Kalman Filter [66].

A discrete estimator called singular or generalized dynamic estimator is now presented: Consider the following dynamic mass balances:

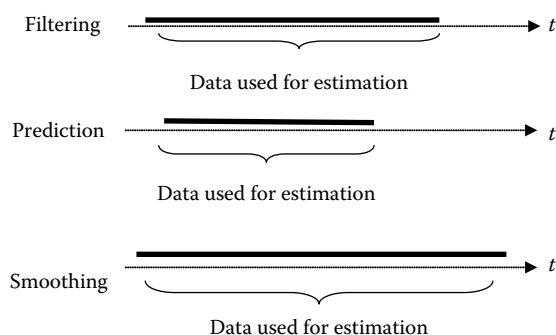


FIG. 22.6
Different types of estimation.

$$B_R \frac{dw_R}{dt} = A_R f_R \quad (22.60)$$

$$C_R f_R = 0 \quad (22.61)$$

where w_R is an inventory variable. The discrete version of these equations is

$$B_R(w_{R,i+1} - w_{R,i}) = A_R f_{R,i+1} \quad (22.62)$$

$$C_R f_{R,i+1} = 0 \quad (22.63)$$

which can be expressed as follows:

$$E_R z_{R,i+1} + G_R z_{R,i} = 0 \quad (22.64)$$

where

$$G_R = \begin{bmatrix} 0 & B_R \\ 0 & 0 \end{bmatrix} \quad (22.65)$$

Unfortunately, E_R is usually not square and therefore singular. Therefore, the discrete Kalman filter cannot be applied. Instead, all estimates for all times up to t_N are reconciled at the same time using the constraint given by (22.64), that is,

$$\left. \begin{array}{l} \text{Min} \sum_{k=0}^N [z_{R,k} - z_{R,k}^+]^T Q_R^{-1} [z_{R,k} - z_{R,k}^+] \\ \text{such that} \\ E_R z_{R,k+1} + G_R z_{R,k} = 0 \quad \forall k = 0, N-1 \end{array} \right\} \quad (22.66)$$

Darouach and Zasadzinski [67] developed recursive formulas that would allow the solution to these models. Rollins and Devanathan [20] presented an alternative methodology that is not as computationally intensive and allows one to trade computational speed for estimation accuracy.

Finally, a smoothing approach based on polynomial representations and integration of the differential equations can be used [68]. Consider the following s -order polynomial representation of f_R and w_R :

$$f_R = \sum_{k=0}^s a_k^R t^k \quad (22.67)$$

$$w_R = w_{R0} + \sum_{k=0}^s w_{k+1}^R t^{k+1} \quad (22.68)$$

therefore,

$$\int_0^t f_R(x) dx = \sum_{k=0}^s \frac{a_k^R}{k+1} t^{k+1} \quad (22.69)$$

Thus, Equation 22.60 is equivalent to

$$B_R [w_R - w_{R0}] = B_R \sum_{k=0}^s w_{k+1}^R t^{k+1} = A_R \sum_{k=0}^s \frac{a_k^R}{k+1} t^{k+1} \quad (22.70)$$

Since this equation is valid for all t , then

$$B_R w_{k+1}^R = \frac{A_R a_k^R}{k+1}, \quad k = 0, \dots, s \quad (22.71)$$

In turn, (22.61) is equivalent to the following set of equations:

$$C_R a_k^R = 0, \quad k = 0, \dots, s \quad (22.72)$$

For $k+1$ measurements, the dynamic data reconciliation problem assumes the following form:

$$\left. \begin{array}{l} \text{Min} \sum_{k=1}^N [z_{R,k} - z_{R,k}^+]^T Q_R^{-1} [z_{R,k} - z_{R,k}^+] \\ \text{such that} \\ B_R (w_{R,k} - w_{R,0}) = A_R \int_0^t f_R(x) dx \quad (k = 0, \dots, N) \\ C_R f_{R,k} = 0 \quad (k = 0, \dots, N) \end{array} \right\} \quad (22.73)$$

Typical results are shown in Figure 22.7.

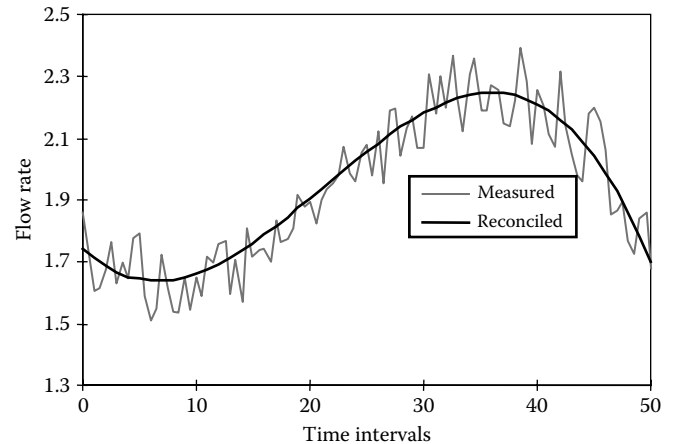


FIG. 22.7

Typical results obtained using integral dynamic data reconciliation.

TABLE 22.6*Data Reconciliation Packages (in Alphabetical Order)*

<i>Package</i>	<i>Nature</i>	<i>Offered by</i>
Aspen Advisor	Commercial	Aspentech (United States)
Datacon	Commercial	Simulation Sciences (United States)
Interactive Online Optimization	Academic	Louisiana State University (United States)
Massbal	Commercial	Hyprotech (Canada)
Production Balance	Commercial	Honeywell (United States)
Recon	Commercial	Chemplant Technologies (Czech Republic)
Reconciler	Commercial	Resolution Integration Solutions (United States)
Sigmafine	Commercial	OSI (United States)
Vali	Commercial	Belsim (Belgium)

Although there is a lot of work that has been performed in the field of dynamic data, its implementation in industrial sites is incipient. Several approaches have not been described in this unit that are of practical value. One good example is the use of orthogonal collocation [69] and the use of differential algebraic solvers [48,70].

Software vendors are still considering dynamic data reconciliation as too computationally intensive and logistically an effort in programming that is not worthwhile. Ultimately, it will be the pressure from the practitioners who will be not satisfied with the steady-state-based approach that will force the commercial implementation of these techniques.

AVAILABLE SOFTWARE

Several software packages exist that are devoted directly to data reconciliation or have data reconciliation embedded as functionality within other programs, and typically yield accounting or online optimization. Table 22.6 depicts some of these available programs.

Almost all packages make use of the weighted least squares and a preprocessing that allows the determination of redundant measurements. However, some use alternative objectives: Interactive Online Optimization (Louisiana State University) allows the use of the combined probability distribution and the Lorentzian objective function. In turn, Production Balance (Honeywell) uses ridge regression, which is a technique that avoids classification by treating unmeasured variables as measured with a high variance. This package has also some alternative classification methods.

Many packages report the precision of the estimators (reconciled values), and some perform studies of interaction between measurements and percent contributions of each measurement to the estimator (Datacon, Vali). A variety of packages also allow considering bounds on system variables.

The incorporation and enhancement of multiple gross error handling seems to be the next step for all commercial codes and many already have it. Indeed, Datacon (Simsci) and Sigmafine (OSI) perform serial elimination based on

the MT but Datacon also allows the user to use principal components, Massbal (Hyprotech) performs serial compensation using GLR, and Vali (Belsim) will shortly add serial elimination.

Several packages are capable of performing material, component, and energy data reconciliation. Among those that perform energy balance reconciliation, some are equipped with very powerful property prediction methods (even liquid vapor equilibrium calculations and handling of pseudocomponents), while others rely on limited correlation methods. This feature should be of concern at usage time because the errors included in the model due to inaccurate correlations propagate to the estimators obtained. Finally, some are able to handle pseudo-components (Datacon, Vali) and even reconcile them (Vali). Finally, many of these packages function together with a production accounting software.

References

1. Johnston L. P. M. and M. A. Kramer, Maximum likelihood data rectification: Steady state systems, *AIChE J.*, 41, 11, 1995.
2. Crowe C. M., Formulation of linear data reconciliation using information theory, *Comp. Chem. Eng.*, 51(12), 3359–3366, 1996.
3. Madron F., *Process Plant Performance*, Ellis Horwood, Chichester, U.K., 1992.
4. Mah R. S. H. *Chemical Process Structures and Information Flows*, Butterworths, London, U.K., 1990.
5. Narasimhan S. and C. Jordache, *Data Reconciliation and Gross Error Detection. An Intelligent Use of Process Data*, Gulf Publishing Company, Houston, TX, 2000.
6. Sánchez M. and J. Romagnoli, *Data Processing and Reconciliation for Chemical Process Operations*, Academic Press, San Diego, CA, 2000.
7. Ragot J. and D. Maquin, Reformulation of data reconciliation problem with unknown-but-bounded errors, *Ind. Eng. Chem. Res.*, 43, 1530–1536, 2004.
8. Narasimhan S. and P. Harikumar, A method to incorporate bounds in data reconciliation and gross error detection—I. The bounded data reconciliation problem, *Comp. Chem. Eng.*, 17(II), 1115–1120, 1993.
9. Kalman R. E., New approach to linear filtering and prediction problems, *J. Basic Eng.*, ASME, 82D, 35–45, 1960.

10. Stanley G. M. and R. H. S. Mah, Observability and redundancy in process data estimation, *Chem. Eng. Sci.*, 36, 259–272, 1981.
11. Maquin D., M. Luong, and J. Ragot, Observability analysis and sensor placement, *Proceedings of Safe Process '94 IFAC/IMACS Symposium on Fault Detection, Supervision and Safety for Technical Process*, June 13–15, Espoo, Finland, 1994.
12. Maquin D., M. Luong, and J. Paris, Dependability and analytical redundancy, *IFAC Symposium on On-Line Fault Detection in the Chemical Process Industries*, Newcastle, U.K., 1995.
13. Bagajewicz M. and M. Sánchez, Design and upgrade of non-redundant and redundant linear sensor networks, *AIChE J.*, 45(9), 1927–1939, 1999.
14. Crowe C. M., Y. A. García Campos, and A. Hrymak, Reconciliation of process flow rates by matrix projection. Part I: Linear case, *AIChE J.*, 29, 881–888, 1983.
15. Swartz C. L. E., *Data Reconciliation for Generalized Flowsheet Applications*, American Chemical Society—National Meeting, Dallas, TX, 1989.
16. Sánchez M. and J. Romagnoli, Use of orthogonal transformations in data classification—Reconciliation, *Comp. Chem. Eng.*, 20, 483–493, 1996.
17. Kuehn D. R. and H. Davidson, Computer control. II. Mathematics of control, *Chem. Eng. Prog.*, 57, 44, 1961.
18. Bagajewicz M. and Q. Jiang, Comparison of steady state and integral dynamic data reconciliation, *Comp. Chem. Eng.*, 24(11), 2367–2518, 2000.
19. Bagajewicz M. and M. Gonzales, Is the practice of using unsteady data to perform steady state reconciliation correct? *AIChE Spring Meeting*, Houston, TX, April 2001.
20. Rollins D. K. and S. Devanathan, Unbiased estimation in dynamic data reconciliation, *AIChE J.*, 39, 8, 1993.
21. Almsy G. A., Principles of dynamic balancing, *AIChE J.*, 36, 1321–1330, 1990.
22. Cao S. and R. R. Rhinehart, An efficient method for on-line identification of steady state, *J. Proc. Cont.*, 5(6), 363–374, 1995.
23. Cao S. and R. R. Rhinehart, Critical values for a steady-state identifier, *J. Process Control*, 7, 149, 1997.
24. Brown P. R. and R. R. Rhinehart, Automated steady-state identification in multivariable systems, *Hydrocarbon Process.*, 79, 79–83, 2000.
25. Narasimhan S., R. S. H. Mah, A. C. Tamhane, J. W. Woodward, and J. C. Hale, A composite statistical test for detecting changes of steady states, *AIChE J.*, 32(9), 1409–1418, 1986.
26. Jiang T., B. Chen, X. He, and P. Stuart, Application of steady-state detection method based on wavelet transform, *Comp. Chem. Eng.*, 27, 569–578, 2003.
27. Bhat S. A. and D. N. Saraf, Steady-state identification, gross error detection, and data reconciliation for industrial process units, *Ind. Eng. Chem. Res.*, 43, 4323–4336, 2004.
28. Bagajewicz M., *Smart Process Plants: Software and Hardware for Accurate Data and Profitable Operations* (ISBN:978-0-07-160471), McGraw Hill, New York, 2010.
29. Almsy G. A. and R. S. H. Mah, Estimation of measurement error variances from process data, *Ind. Eng. Chem. Process Des. Dev.*, 23, 779, 1984.
30. Darouach M., R. Ragot, M. Zasadzinski, and G. Krzakala, Maximum likelihood estimator of measurement error variances in data reconciliation, *IFAC. AIPAC Symposium 2*, pp. 135–139, 1989.
31. Keller J. Y., M. Zasadzinski, and M. Darouach, Analytical estimator of measurement error variances in data reconciliation, *Comp. Chem. Eng.*, 16, 185, 1992.
32. Chen J., A. Bandoni, and J. A. Romagnoli, Robust estimation of measurement error variance/covariance from process sampling data, *Comp. Chem. Eng.*, 21(6), 593–600, 1997.
33. Mah R. S. H. and A. C. Tamhane, Detection of gross errors in process data, *AIChE J.*, 28, 828, 1982.
34. Sidák Z., Rectangular confidence regions for the means of multivariate normal distributions, *J. Am. Stat. Assoc.*, 62, 626–633, 1967.
35. Rollins D. K. and J. F. Davis, Unbiased estimations of gross errors in process measurements, *AIChE J.*, 38, 563–572, 1992.
36. Rollins D. K., Y. Cheng, and S. Devanathan, Intelligent selection of hypothesis tests to enhance gross error identification, *Comp. Chem. Eng.*, 20(5), 517–530, 1996.
37. Rollins D. K. and J. F. Davis, Gross error detection when variance-covariance matrices are unknown, *AIChE J.*, 39(8), 1335–1341, 1993.
38. Crowe C. M., Observability and redundancy of process data for steady state reconciliation, *Chem. Eng. Sci.*, 44, 2909–2917, 1989.
39. Narasimhan S. and R. S. H. Mah, Generalized likelihood ratio method for gross error identification, *AIChE J.*, 33, 1514–1521, 1987.
40. Tong H. and C. M. Crowe, Detection of gross errors in data reconciliation using principal component analysis, *AIChE J.*, 41, 1712–1722, 1995.
41. Chen V. C. P., M. Melendez, and D. K. Rollins, The problem of too much Power in detecting biases in real chemical processes, *ISA Trans.*, 37, 329–336, 1998.
42. Iordache C., R. Mah, and A. Tamhane, Performance studies of the measurement test for detection of gross errors in process data, *AIChE J.*, 31, 1187, 1985.
43. Bagajewicz M. and Q. Jiang, Gross error modeling and detection in plant linear dynamic reconciliation, *Comp. Chem. Eng.*, 22(12), 1789–1810, 1998.
44. Bagajewicz M., Q. Jiang, and M. Sánchez, Removing singularities and assessing uncertainties in two efficient gross error collective compensation methods, *Chem. Eng. Commun.*, 178, 1–20, 2000.
45. Jiang Q. and M. Bagajewicz, On a strategy of serial identification with collective compensation for multiple gross error estimation in linear data reconciliation, *Ind. Eng. Chem. Res.*, 38(5), 2119–2128, 1999.
46. Tjoa I. B. and L. T. Biegler, Simultaneous strategies for data reconciliation and gross error detection of nonlinear systems, *Computers Chem. Eng.*, 15(10), 679–690, 1991.
47. Mah R. S. H., Letter to the editor, *Comp. Chem. Eng.*, 21(9), 1069, 1997.
48. Albuquerque J. S. and L. T. Biegler, Data reconciliation and gross-error detection for dynamic systems, *AIChE J.*, 42(10), 2841, 1996.
49. Soderstrom T. A., D. M. Himmelblau, and T. F. Edgar, A mixed integer optimization approach for simultaneous data reconciliation and identification of measurement bias, *Control Eng. Pract.*, 9(8), 869–876, 2001.
50. Arora N. and L. T. Biegler, Redescending estimators for data reconciliation and parameter estimation, *Comp. Chem. Eng.*, 25, 1585–1599, 2001.
51. Wang D. and J. A. Romagnoli, A framework for robust data reconciliation based on a generalized objective function, *Ind. Eng. Chem. Res.*, 42(13), 3075–3084, 2003.
52. Ozyurt D. B. and R. W. Pike, Theory and practice of simultaneous data reconciliation and gross error detection for chemical processes, *Comp. Chem. Eng.*, 28, 381–402, 2004.

53. Morad K., B. R. Young, and W. Y. Svrcek, Rectification of plant measurements using a statistical framework, *Comp. Chem. Eng.*, 29, 919–940, 2005.
54. Alhaj-Dibo M., D. Maquin, and J. Ragot, Data reconciliation: A robust approach using a contaminated distribution, *Control Eng. Pract.*, 16(2), 159–170, 2008.
55. Bakshi B. R., M. N. Nounou, P. K. Goel, and X. Shen, Multiscale Bayesian rectification of data from linear steady-state and dynamic systems without accurate models, *Ind. Eng. Chem. Res.*, 40(1), 261–274, 2001.
56. Chen W., B. R. Bakshi, P. K. Goel, and S. Ungarala, Bayesian estimation via sequential Monte Carlo sampling: Unconstrained nonlinear dynamic systems, *Ind. Eng. Chem. Res.*, 43, 4012–4025, 2004.
57. Miller R. W., *Flow Measurement Engineering Handbook*, McGraw Hill, New York, 1996.
58. Bagajewicz M., On the definition of software accuracy in redundant measurement systems, *AIChE J.*, 51(4), 1201–1206, 2005.
59. Bagajewicz M. and D. T. Nguyen, Stochastic-based accuracy of data reconciliation estimators for linear systems, *Comp. Chem. Eng.*, 32(6), 1257–1269, 2008.
60. Romagnoli J. A., On data reconciliation-constraints processing and treatment of bias, *Chem. Eng. Sci.*, 38, 1107–1117, 1983.
61. Joris P. and B. Kalitventzeff, Process measurement analysis and validation, *Proceedings of CEF'87: Use of Computers and Chemical Engineering*, Italy, pp. 41–46, 1987.
62. Maquin D., G. Bloch, and J. Ragot, Data reconciliation of measurements, *Diagnostic et surete de fonctionnement*, 1, 145–181, 1991.
63. Meyer M., B. Koehert, and M. Enjalbert, Data reconciliation on multicomponent network processes, *Comp. Chem. Eng.*, 17, 807–817, 1993.
64. Crowe C. M., Data reconciliation: Progress and challenges, *J. Process Control*, 6, 89–98, 1996.
65. Stanley G. M. and R. H. S. Mah, Estimation of flows and temperatures in process networks, *AIChE J.*, 23(5), 642, 1977.
66. Karjala, T. W. and D. M. Himmelblau, Dynamic rectification of data via recurrent neural nets and the extended kalman filter, *AIChE J.*, 42, 2225, 1996.
67. Darouach M. and M. Zasadzinski, Data reconciliation in generalized linear dynamic systems, *AIChE J.*, 37(2), 193, 1991.
68. Bagajewicz, M. and Q. Jiang, An integral approach to dynamic data reconciliation, *AIChE J.*, 43, 2546, 1997.
69. Liebman M. J., T. F. Edgar, and L. S. Lasdon, Efficient data reconciliation and estimation for dynamic process using nonlinear programming techniques, *Chem. Eng. Sci.*, 16(10/11), 963, 1992.
70. Albuquerque J. S. and L. T. Biegler, Decomposition algorithms for on-line estimation with nonlinear DAE models, *Comp. Chem. Eng.*, 19, 1031, 1995.